Case study - MLOps engineer

Context

For several years now, the sales performance team have been using a **speech-to-text service** to transcribe the audio of telephone calls made by sales staff. This enables them to better understand what's working well and what's not.

At the end of the year, the contract came to an end and the costs seemed to be quite high. The software engineers on the Sales side knew that an **ML Platform team** existed, and that they would be able to set up an **in-house speech-to-text solution**.

Use case

So we're offering you the chance to self-host the Whisper open source model and make a first version for the teams working on its integration. The aim is to clearly define all the areas of exploration and show the team a first MVP.

- The MVP would have to be functional and scalable for **100 sales**. It's worth noting that the solution should be scalable for all sales (x10 or so), and that we'll probably need advice on how to improve the solution's scaling.
- This MVP will have to be integrated by software engineers for testing, so the **API contract** will have to be clear (take an audio format as input and text as output).

As a manager, I'd like you to give us a framework for your work and be able to present us with the main trade-offs and end-to-end work on this issue.

Here are a few suggestions :

- What are the main challenges to be addressed ?
- What would architecture design be ?
- How to build a scalable inference code (python / libs) ?
- How to containerize your code (docker) and how to serve it scalably ?
- How to build a scalable infrastructure with **terraform** ? What is the deployment process ?
- What infrastructure issues need to be addressed ?
- How to monitor the performance of the solution ?
- What are the limitations and possible improvements for your solution ?

Instructions

***What we want (must have)** :*

- A working document / tech scopings to explain your technical choices and your recommendations (a README can be ok).
- A structured git repository with your technical explorations.
- Ideally, something that would allow you to show your technical expertise.
- An indication of how much time you spent on this case. The evaluation will take into account the work / time spent trade-off.
- Don’t hesitate to prioritize your work, and explain what you wanted to add but didn’t have the time to do given the timeframe you had to complete the case.

What we prefer :

- The case study is deliberately broad and the candidate is not expected to answer all the points.
- One well explored subject is preferable to several poorly worked out ones : we prefer to have a demo from A to Z on the demo/staging-part, rather than an attempt to cover everything at the expense of depth on every topic.

What is at your discretion :

- The presentation (slides, document, code, markdown, no support) is at the candidate's discretion.

What we absolutely don’t want :

- We absolutely don’t want a high-level presentation that would not allow us to proxy your technical expertise.
- We absolutely don’t want a proposal that doesn't match our existing stack. For example, use of GCP or Gitlab CI/CD.

Don’t hesitate to ask the ML platform team (ml-platform@doctolib.com) if you have any questions related to the case, our stack, our data model...